Offensive Language Identification in Transliterated and Code-Mixed Bangla

Md Nishat Raihan[1], Umma Hani Tanmoy[1], Anika Binte Islam[1], Kai North[1]
Tharindu Ranasinghe[2], Antonios Anastasopoulos[1], Marcos Zampieri[1]
[1]George Mason University, USA
[2]Aston University, UK
mraihan2@gmu.edu

Abstract

Identifying offensive content in social media is vital for creating safe online communities. Several recent studies have addressed this problem by creating datasets for various languages. In this paper, we explore offensive language identification in texts with transliterations and code-mixing, linguistic phenomena common in multilingual societies, and a known challenge for NLP systems. We introduce TB-OLID, a transliterated Bangla offensive language dataset containing 5,000 manually annotated comments. We train and fine-tune machine learning models on TB-OLID, and we evaluate their results on this dataset. Our results show that English pre-trained transformer-based models, such as fBERT and HateBERT achieve the best performance on this dataset.

1 Introduction

As the popularity of social media continues to grow, the spread of offensive content in these platforms has increased substantially, motivating companies to invest heavily in content moderation strategies and robust models to detect offensive content. We have observed a growing interest in this topic, evidenced by popular shared tasks at SemEval (Basile et al., 2019) and other venues. Apart from a few notable exceptions (Mandl et al., 2020), most of the work on this topic has not addressed the question of transliteration and code-mixing, two common phenomena in social media.

Code-mixing is the phenomenon of embedding linguistic units such as phrases, words, or morphemes of one language into another language (Myers-Scotton, 1997; Muysken et al., 2000). Code-mixed texts often feature transliterations where speakers use an alternative script to the language's official or standard script by mapping from one writing system (e.g., Hindi and its original Devanagari script) to another one (e.g., Latin transliteration of Hindi) based on phonetic similarity. Transliterated texts are widely used in social media platforms as transliteration allows users to write in their native language using a script that may not be supported by the platform and/or using Latin-based default keyboards. Furthermore, the use of transliteration also allows users to easily switch between languages with otherwise different scripts (e.g., English and Hindi). As discussed in a recent survey (Winata et al., 2022), however, processing code-mixing datasets is a challenge that hinders performance in a variety of NLP tasks, thus deserving special attention.

Code-mixing and transliteration are common in various languages, including Bangla (Das and Gambäck, 2015; Jamatia et al., 2015). Related work on Bangla offensive language identification (Wadud et al., 2021), however, has mostly focused on standard Bangla script. As such, the performance of offensive language identification models on code-mixing and transliterated Bangla remains largely unexplored. To address this shortcoming, we create TB-OLID, a manually annotated transliterated Bangla offensive language dataset. TB-OLID was annotated following the popular OLID hierarchical taxonomy (Zampieri et al., 2019a), allowing cross-lingual experiments. To the best of our knowledge, the dataset is the first of its kind for Bangla, opening exciting new avenues for future research.

The main contributions of this paper are as follows:

1. We introduce TB-OLID, an offensive language identification corpus containing 5,000 Facebook comments.[1]
2. We provide a comparative analysis of various machine learning models trained or fine-tuned on TB-OLID.

WARNING: This paper contains examples that are offensive in nature.

[1]https://github.com/LanguageTechnologyLab/TB-OLID

Proceedings of the First Workshop on Bangla Language Processing (BLP-2023), pages 1–6

Comment	C|T	O|N	I|G|U
BN: *Tui to 1 ta rastar chele mother chod.* EN: You are a motherfucking street vagabond	**C**	**O**	**I**
BN: *O to Manush na sokun* EN: He/She is not a person but a vulture	**T**	**O**	**I**
BN: *R kichudin por kanglu ra Uth ar mut diye cha banabe* EN: After some days, the barbarians will make tea with camel piss	**T**	**O**	**G**
BN: *Dhoren r dog ar baccha gulo ke gono dholai diye pongu kore den* EN: Capture these son of bitches and beat them to their death	**C**	**O**	**G**
BN: *Pagole kina bole chagole kina khai* EN: A mad man and an animal have no difference	**T**	**O**	**U**
BN: *Ami kintu parlam na hojom korte* EN: I cannot fathom it anymore	**T**	**N**	

Table 1: Examples from TB-OLID in Bangla along with an English translation. The labels included are C (transliterated code-mixed), T (transliterated Bangla), O (offensive), N (not-offensive), I (offensive posts targeted at an individual), G (offensive posts targeted at a group), and U (untargeted offensive posts).

2 Data

Data Collection We collect data from Facebook, the most popular social media platform in Bangladesh. We compile a list of the most popular Facebook pages in Bangladesh using Fanpage Karma[2] and scraped comments from each of the top 100 most followed Facebook pages using the publicly available Facebook scraper tool.[3] This results in an initial corpus of over 100,000 comments. We exclude all comments not written with non-Latin script. We search the corpus using keywords for transliterated hate speech and offensive language. We select keywords from the list of 175 offensive Bangla terms by Karim et al. (2021). As the dataset by Karim et al. (2020) contains standard Bangla, we convert keywords into transliterated Bangla using the Indic-transliteration tool.[4] Using these keywords we randomly select a set of 5,000 comments for annotation.

Annotation Guidelines We prepare the TB-OLID annotation guidelines containing labels and examples. The first step is to label whether a comment is transliterated Bangla or transliterated code-mixed Bangla. If the comment contains at least one English word along with other Bangla transliterated words, we consider it as transliterated code-mixed. Next, we consider the offensive vs. non-offensive distinction and, in the case of offensive posts, its target or lack thereof. Table 1 presents six annotated instances included in TB-OLID.

We adopt the guidelines introduced by the popular OLID annotation taxonomy (Zampieri et al., 2019a) used in the OffensEval shared task (Zampieri et al., 2019b) and replicated in multiple other datasets in languages such as Danish (Sigurbergsson and Derczynski, 2020), Greek (Pitenis et al., 2020), Marathi (Gaikwad et al., 2021; Zampieri et al., 2022), Portuguese (Sigurbergsson and Derczynski, 2020), Sinhala (Ranasinghe et al., 2022) and Turkish (Çöltekin, 2020). We choose OLID due to the flexibility provided by its three-level hierarchical taxonomy that allows us to model different types of offensive and abusive content (e.g., hate speech, cyberbulling, etc.) using a single taxonomy. OLID's taxonomy considers whether an instance is offensive (level A), whether an offensive post is targeted or untargeted (level B), and what is the target of an offensive post (level C). As the second level of the TB-OLID annotation we consider OLID level A as follows.

- **Offensive:** Comments that contain any form of non-acceptable language or a targeted offense, including insults, threats, and posts containing profane language
- **Non-offensive:** Comments that do not contain any offensive language

Finally, the third level of the TB-OLID annotation merges OLIDs level B and C. We label whether a post is untargeted or, when targeted, whether it is labeled at an individual or a group as follows:

- **Individual:** Comments targeting any individ-

[2] https://www.fanpagekarma.com/
[3] https://github.com/kevinzg/facebook-scraper
[4] https://github.com/sanskrit-coders/indic_transliteration_py

ual, such as mentioning a person with his/her name, unnamed participants, or famous personality.

- **Group:** Comments targeting any group of people of common characteristics, religion, gender, etc.
- **Untargeted:** Comments containing unacceptably strong language or profanities that are not targeted.

Ensuring Annotation Quality Three annotators working on this project are tasked to annotate TB-OLID. They are PhD students in Computing aged 22-28 years old, 1 male and 2 female, all native speakers of Bangla and fluent speakers of English. The first step of the annotation process involves a pilot annotation study, where 300 comments are assigned to all three annotators to calculate initial agreement and refine the annotation guidelines according to their feedback. After this pilot experiment, we annotate an additional 4,700 Facebook comments totaling 5,000 instances which are subsequently split into 4,000 and 1,000 instances for training and testing, respectively. The instances in TB-OLID are annotated by at least two annotators, with the third one serving as adjudicator. We calculate pairwise inter-annotator agreement on 1,000 instances using Cohen's Kappa, and we report Cohen's Kappa score of 0.77 and 0.72 for levels 1 (code-mixed vs. transliterated) and 2 (offensive vs. non-offensive), which is generally considered substantial agreement. We report Cohen's Kappa score of 0.66 on level 3, considered moderate agreement.

Dataset Statistics We calculated the frequency of each label in the dataset namely code-mixed and transliterated, offensive and non-offensive, targeted and untargeted, and target types in the dataset. The dataset statistics are presented in Table 2.

Level	Label	Instances	Percentage
1	T	2,959	59.18%
	C	2,041	41.82%
2	O	2,381	47.62%
	N	2,619	52.38%
3	I	1,192	23.84%
	G	954	19.08%
	U	235	4.70%

Table 2: TB-OLID per level and per class statistics. Percentage calculated considering the total number of instances in the dataset (5,000).

Finally, we run an analysis of the code-mixed data using ad-hoc Python scripts. We observe that English is by far the most common language included in the code-mixed instances mixed with Bangla followed by Hindi. We report that 38.42% of all tokens in the code-mixed (C) class are English.

3 Baselines and Models

Baselines We report the results of three baselines: (1) Google's Perspective API[5], a free API developed to detect offensive comments widely used as a baseline in this task (Kaati et al., 2022; Fortuna et al., 2020); (2) prompting GPT 3.5 turbo providing the model with TB-OLID's annotation guidelines; and (3) a majority class baseline. Due to the API's limitations, Perspective API was used only for offensive language identification and not for for target classification.

General Models We experiment with pre-trained language models fine-tuned on TB-OLID. As our dataset is transliterated Bangla and contains English code-mixed, we experiment with BERT (Devlin et al., 2019), roBERTa (Liu et al., 2020) which are trained on English, and Bangla-BERT (Kowsher et al., 2022), which is trained on Bangla. We also use cross-lingual models such as mBERT (Devlin et al., 2019) and xlm-roBERTa (Conneau et al., 2020) which are trained in multiple languages.

Task-specific Models We also experiment with task-specific fined-tuned models like HateBERT (Caselli et al., 2021), and fBERT (Sarkar et al., 2021). These models were also further fined-tuned on TB-OLID.

4 Results and Discussion

We use F1-score to evaluate the performance of all models. The training and test sets are obtained by the aforementioned 80-20 random split on the entire TB-OLID dataset. We further subdivide the test set into transliterated code-mixed (C), transliterated (T), and all instances. We present results for offensive text classification (offensive vs. non-offensive) in Table 3.

We observe that the standard BERT model performs well over the baselines, whereas the Bangla-BERT model performs less well. BERT achieves F1-score of 0.71, whereas Bangla-BERT obtains F1-score of 0.42. We believe this is due to the

[5] https://perspectiveapi.com/

Model	C	T	All
fBERT	0.73	0.70	0.72
HateBERT	0.74	0.69	0.72
BERT	0.73	0.68	0.71
m-BERT	0.70	0.68	0.69
GPT 3.5	0.65	0.64	0.64
Majority Class Baseline	0.57	0.57	0.57
Perspective API	0.53	0.50	0.51
Bangla-BERT	0.42	0.42	0.42
xlm-roBERTa	0.40	0.41	0.41
roBERTa	0.41	0.41	0.41

Table 3: Offensive Language Identification - F1-score of all models trained and/or fine-tuned on TB-OLID. We report results on the transliterated code-mixed (C), transliterated (T), and All test set. Baselines in italics.

fact that many instances in the dataset are in Latin script, which means that BanglaBERT frequently struggles with out-of-vocabulary tokens. The low performance of Bangla-BERT in this task requires further examination. Models pre-trained specifically on offensive language identification perform very well with fBERT and Hate-BERT coming out on top, both with an F1 score of 0.72. Finally, we observe that the top-5 performing models perform better on the code-mixing data compared to transliterated data. This is likely due to the heavy presence of English words in the code-mixing data where we observe the presence of 38% of English words.

Finally, Table 4 presents the results of target type classification (individual, group, or untargeted).

Model	C	T	All
HateBERT	0.69	0.66	0.68
m-BERT	0.72	0.64	0.67
BERT	0.72	0.64	0.67
fBERT	0.66	0.64	0.65
roBERTa	0.73	0.60	0.65
GPT 3.5	0.39	0.46	0.43
Majority Class Baseline	0.48	0.63	0.55
xlm-roBERTa	0.61	0.51	0.55
Bangla-BERT	0.59	0.47	0.51

Table 4: Target Classification - F1-score of all models trained and/or fine-tuned on TB-OLID. We report results on the transliterated code-mixed (C), transliterated (T), and All test sets. Baselines in italics.

Overall, target classification is a more challenging task than offensive language identification due to the presence of three classes instead of two. Therefore, all results are substantially lower for this task. HateBERT performs better than all other models with an F1 score of 0.68. roBERTa achieved more competitive performance for target classification than for offensive language identification whereas Bangla-BERT did not perform well in both tasks. Finally, similar to the previous task, the best-performing models achieved higher F1 scores on the code-mixed data than on the transliterated data.

One key observation is that the transformer-based models do not perform very well, since most of them are not pre-trained on transliterated Bangla. Among the models that we experiment with, only xlm-roBERTa is pre-trained with a comparatively small set of Romanized Bangla. However, the lack of any standard rules for spelling in transliterated Bangla makes TB-OLID very challenging.

5 Conclusion and Future Work

In this work, we introduced TB-OLID, a transliterated Bangla offensive language dataset containing 5,000 instances retrieved from Facebook. Three native speakers of Bangla have annotated the dataset with respect to the presence of code-mixing, the presence of offensive language, and its target according to the OLID taxonomy. TB-OLID opens exciting new avenues for research on offensive language identification in Bangla.

We performed experiments with multiple models such as general monolingual models like BERT (Devlin et al., 2019), roBERTa (Liu et al., 2020) and Bangla-BERT (Kowsher et al., 2022); cross-lingual models like mBERT (Devlin et al., 2019) and xlm-roBERTa (Conneau et al., 2020); and models fine-tuned for offensive language identification like HateBERT (Caselli et al., 2021), and fBERT (Sarkar et al., 2021)). The best results were obtained by the task-specific models.

In future work, we would like to extend the TB-OLID dataset and annotate the offense type (e.g., religious offense, political offense, etc.). This would help us identify the common targets in various platforms. Furthermore, we would like to pre-train and fine-tune a Bangla transliterated BERT model to see how it performs on TB-OLID. Finally, in future work, we would like to evaluate the performance of other recently released large language models (LLMs) (e.g., GPT 4.0, Llama 2) on TB-OLID. The first baseline results using GPT 3.5 indicate that general-purpose LLMs still struggle with the transliterated and code-mixed content presented in TB-OLID.

Acknowledgments

We thank the anonymous workshop reviewers for their insightful feedback. Antonios Anastasopoulos is generously supported by NSF award IIS-2125466.

Ethics Statement

The generation and annotation procedure of TB-OLID adheres to the ACL Ethics Policy and seeks to make a valuable contribution to the realm of online safety. The technology in question possesses the potential to serve as a beneficial instrument for the moderation of online content, thereby facilitating the creation of safer digital environments. However, it is imperative to exercise caution and implement stringent regulations to prevent its potential misuse for purposes such as monitoring or censorship.

References

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of SemEval*.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. In *Proceedings of WOAH*.

Çağrı Çöltekin. 2020. A Corpus of Turkish Offensive Language on Social Media. In *Proceedings of LREC*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Amitava Das and Björn Gambäck. 2015. Code-mixing in social media text: The last language identification frontier? *Revue TAL - Association pour le Traitement Automatique des Langues (ATALA)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive, or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of LREC*.

Saurabh Sampatrao Gaikwad, Tharindu Ranasinghe, Marcos Zampieri, and Christopher Homan. 2021. Cross-lingual offensive language identification for low resource languages: The case of marathi. In *Proceedings of RANLP*.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Proceedings of RANLP*.

Lisa Kaati, Amendra Shrestha, and Nazar Akrami. 2022. A machine learning approach to identify toxic language in the online space. In *Proceedings of ASONAM*.

Md Rezaul Karim, Bharathi Raja Chakravarthi, John P McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *Proceedings of DSAA*.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *Proceedings of DSAA*.

M Kowsher, Abdullah As Sami, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, Pranab Kumar Dhar, and Takeshi Koshiba. 2022. Bangla-bert: transformer-based efficient model for transfer learning and language understanding. *IEEE Access*, 10:91855–91870.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach. In *Proceedings of ACL*.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive language identification in tamil, malayalam, hindi, english and german. In *Proceedings of FIRE*.

Pieter Muysken et al. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Zesis Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. 2020. Offensive language identification in greek. In *Proceedings of LREC*.

Tharindu Ranasinghe, Isuri Anuradha, Damith Premasiri, Kanishka Silva, Hansi Hettiarachchi, Lasitha Uyangodage, and Marcos Zampieri. 2022. SOLD: Sinhala offensive language dataset. *arXiv preprint arXiv:2212.00851*.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fbert: A neural transformer for identifying offensive content. In *Findings of the ACL*.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive Language and Hate Speech Detection for Danish. In *Proceedings of LREC*.

Md Anwar Hussen Wadud, Md Abdul Hamid, Muhammad Mostafa Monowar, and Atif Alamri. 2021. L-boost: Identifying offensive texts from social media post in bengali. *Ieee Access*, 9:164681–164699.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2022. The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. In *Findings of the ACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of SemEval*.

Marcos Zampieri, Tharindu Ranasinghe, Mrinal Chaudhari, Saurabh Gaikwad, Prajwal Krishna, Mayuresh Nene, and Shrunali Paygude. 2022. Predicting the type and target of offensive social media posts in Marathi. *Social Network Analysis and Mining*, 12(1).